Summarizing Posterior Distributions & Bayesian Model Selection

Prof. Nicholas Zabaras
Center for Informatics and Computational Science
https://cics.nd.edu/
University of Notre Dame
Notre Dame, Indiana, USA

Email: nzabaras@gmail.com
URL: https://www.zabaras.com/

January 19, 2019

Contents

- Summarizing Posterior Distributions – An Introduction, MAP Estimation, Drawbacks of MAP Estimation, MAP Estimation and Reparametrization

- Credible Intervals, HPD Intervals, Bayesian Inference for a Difference in Proportions

- Model Selection and Cross Validation, AIC Information Criterion, Bayesian Model Selection, Bayesian Occam's Razor, Marginal Likelihood, Evidence Approximation, Examples

- Laplace Approximation, Bayesian Information Criterion, Akaike Information Criterion, Effect of the Prior, Empirical Bayes, Bayes Factors and Jeffreys Scale of Evidence, Examples, Jeffreys-Lindley Paradox

- Following closely Chris Bishops' PRML book, Chapters 1 & 2
- Kevin Murphy's, Machine Learning: A probablistic perspective, Chapter 5
- C P Robert, The Bayesian Choice: From Decision-Theoretic Motivations to Compulational Implementation, Springer-Verlag, NY, 2001 (online resource)
- A. Gelman, JB Carlin, HS Stern and DB Rubin, Bayesian Data Analysis, Chapman and Hall CRC Press, 2nd Edition, 2003.
- M Marin and C P Robert, The Bayesian Core, Spring Verlag, 2007 (online resource)
- Bayesian Statistics for Engineering, Online Course at Georgia Tech, B. Vidakovic.

Introduction

- We assume that we computed the posterior of unknown parameters from data.

- Using the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ to summarize everything we know about these variables is at the core of Bayesian statistics.

- We discuss here this approach to statistics in more detail. In particular, we emphasize that point estimates are not the best approach.

- We discuss next some simple quantities that can be derived from $p(\boldsymbol{\theta}|\mathcal{D})$, such as

  - ✓ the posterior mean,
  - ✓ the MAP estimate,
  - ✓ the median, etc.

- These summary statistics (point estimates) are often easier to understand and visualize than the full posterior distribution.

MAP Estimation

- Typically the posterior mean or median is the most appropriate choice for a real-valued quantity, and the vector of posterior marginals is the best choice for a discrete quantity.

- However, *the posterior mode, aka the MAP estimate*, is the most popular choice because it *reduces to an optimization problem*, for which efficient algorithms often exist.

  - ✓ *There are various drawbacks to MAP estimation*, which we discuss below.

- This will provide motivation for a more thoroughly Bayesian approach.

Drawbacks of MAP Estimation

- The most obvious drawback of MAP estimation (and other point estimates such as the posterior mean or median) is *that it does not provide any measure of uncertainty.*

- In many applications, it is important to know how much one can trust a given estimate.

- We will derive such confidence measures from the posterior.

Plugging-in the MAP Estimate Can Overfitt

- In machine learning, we *care more about predictive accuracy than in interpreting the parameters of our models.*

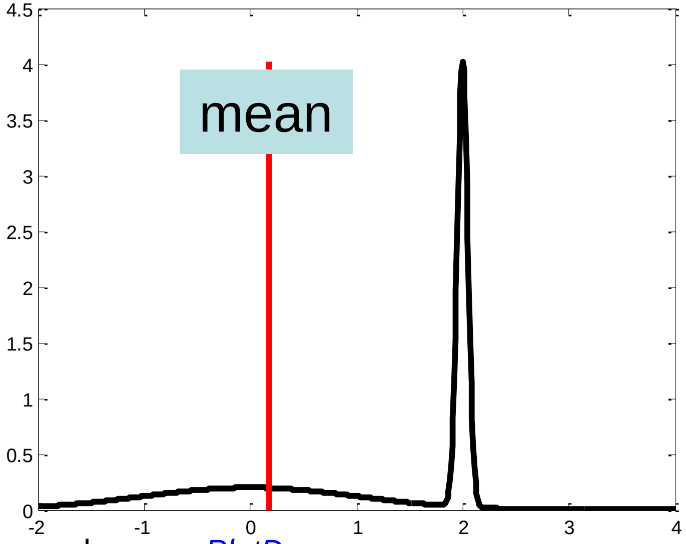
- However, *if we don't model the uncertainty in our parameters, then our predictive distribution will be overconfident*.

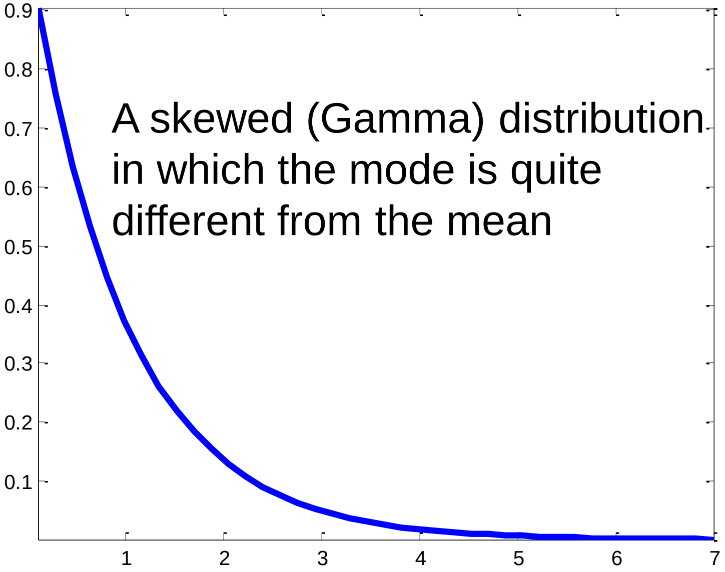
- *Overconfidence in predictions is particularly problematic in situations where we don’t want to take any risk*.

The Mode is an Untypical Estimate

- The mode is usually quite untypical of the distribution, unlike the mean or median (left Fig.)
- The basic problem is that the mean/median take the whole probability mass into account.
- On the example on the right, the mode is 0, but the mean is non-zero. Such *skewed distributions* arise *when inferring variance parameters*, especially in hierarchical models. In such cases the MAP estimate is a very bad estimate.

The mode is very untypical of the distribution. The mean is a better summary of the distribution, since it is near the majority of the probability mass




A skewed (Gamma) distribution in which the mode is quite different from the mean

Run *bimodalDemo* and *gammaPlotDemo* from Kevin Murphys' PMTK

Decision Theory and Loss Functions

- How should we summarize a posterior if the mode is not a good choice?

- The answer is to use decision theory. The basic idea is to specify a loss function, where $L(\theta,\hat{\theta})$ is the loss you incur if the truth is θ and your estimate is $\hat{\theta}$.

- If we use $0-1$ *loss*, $L(\theta,\hat{\theta})=\mathbb{I}(\theta\neq\hat{\theta})$, then the optimal estimate is the *posterior mode*. $0-1$ loss means you only get fully penalized if you make errors.

- For continuous-valued quantities, we often prefer to use *squared error loss*, $L(\theta,\hat{\theta})=(\theta-\hat{\theta})^2$; the optimal estimator is now the *posterior mean*.

- Or we can use a more *robust loss function*, $L(\theta,\hat{\theta})=|\theta-\hat{\theta}|$, which gives rise to the *posterior median*.

MAP Estimation and Reparametrization

- In MAP estimation the result we get depends on the parametrization of the probability distribution. Changing from one representation to another equivalent representation changes the result, which is not very desirable, since the units of measurement are arbitrary (e.g., in measuring distance, we can use cm or in).

- Suppose we compute the posterior for x. If we define $y = f(x)$, the distribution for y is given by

$$p_y(y) = p_x(x)\left|\frac{dx}{dy}\right|$$

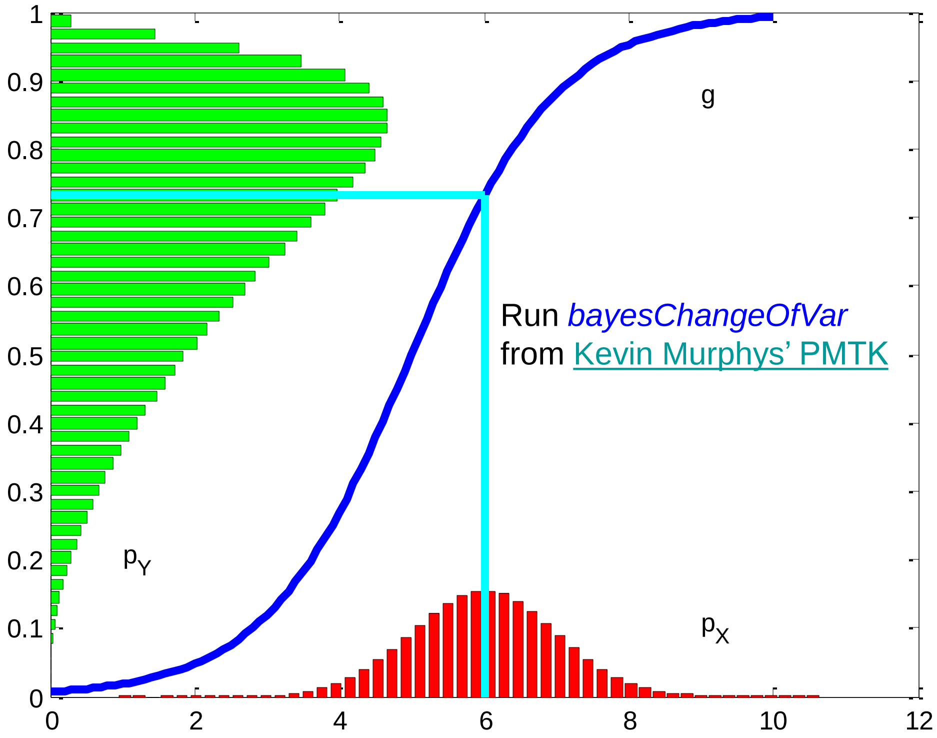
- The Jacobian measures the change in size of a unit volume passed through f.

$$If\ \hat{x} = \max_x p_x(x),\ \hat{y} = \max_y p_y(y) \Rightarrow \hat{y} \neq f(\hat{x})$$

- For example, let $x \sim \mathcal{N}(6,1)$, and $y = f(x) = 1/(1 + \exp(-x + 5))$. We can derive the distribution of y using MC. See the following implementation.

MAP Estimation and Reparametrization

- Transformation of a density under a nonlinear transform. Note how the mode of the transformed distribution is not the transform of the original mode.



Run *bayesChangeOfVar* from [Kevin Murphys' PMTK]

- *The MLE does not suffer from this issue (the likelihood is a function not a probability density).*
- *Bayesian inference also does not suffer from this since the change of measure is taken into account when integrating in the parameter space*.

.

MAP Estimation and Reparametrization

- Consider a Bernoulli likelihood and uniform prior as follows:

$$p(y=1|\mu)=\mu, \, where \; y\in\{0,1\}$$

$$p(\mu)=1, 0\leq\mu\leq 1$$

- Without data, the MAP estimate is the mode of the prior which is anywhere on the interval [0,1]

- *Case 1:* Consider the parametrization

$$\theta=\sqrt{\mu}\Rightarrow|d\mu/d\theta|=2\theta\Rightarrow p_\theta(\theta)=2\theta$$

  - The MAP estimate is $\theta=1$

- *Case 2:* Consider the parametrization:

$$\theta=1-\sqrt{1-\mu}\Rightarrow|d\mu/d\theta|=2(1-\theta)\Rightarrow p_\theta(\theta)=2(1-\theta)$$

  - The MAP estimate is now

$$\theta=0$$

MAP Estimation and Reparametrization

- Using the Fisher information matrix $I(\boldsymbol{\theta})$ associated with the likelihood $p(\boldsymbol{x}|\boldsymbol{\theta})$, a solution to the problem is to optimize the following objective function:

$$\widehat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) |I(\boldsymbol{\theta})|^{-1/2}$$

- This estimate is parameterization independent.

- The optimization problem above is difficult to implement in practice.

- Druilhet, P. and J.-M. Marin (2007). Invariant HPD credible sets and MAP estimators. *Bayesian Analysis 2*(4), 681–692.
- Jermyn, I. (2005). Invariant Bayesian estimation on manifolds. *Annals of Statistics 33*(2), 583–605.

Credible Intervals

- In addition to point estimates, we often want a measure of confidence. A standard measure of confidence in some (scalar) quantity θ is the "width" of its posterior distribution. This can be measured using a $100(1-\alpha)\%$ credible interval, which is a (contiguous) region $C = (\ell, u)$ (standing for lower and upper) which contains $1-\alpha$ of the posterior probability mass, i.e.,

$$C_\alpha(\mathcal{D}) = (\ell, u) : p(\ell \le \theta \le u \mid \mathcal{D}) = 1-\alpha$$

- There may be many such intervals, so we choose one such that *there is $\alpha/2$ mass in each tail; this is called a central interval.*

- Some examples:
  - Gaussian distribution:

Run *quantileDemo*
from Kevin Murphys' PMTK

$$For\ \ p(\theta \mid \mathbf{D}) = \mathcal{N}(0,1),\ (\ell,u) = \big(\Phi(\alpha/2), \Phi(1-\alpha/2)\big) = (-1.96, +1.96)$$

$$For\ \ p(\theta \mid \mathbf{D}) = \mathcal{N}(\mu,\sigma^2),\ (\ell,u) \approx \mu \pm 2\sigma$$

  - Beta prior in a coin example. The posterior is Beta.

Run *betaCreditbleInt*
from Kevin Murphys' PMTK

$$For\ \ p(\theta \mid \mathbf{D}) = \mathcal{Beta}(48,54)\ (47\ H\ in\, 100\ trials),\ (\ell,u) = \big(0.3749, 0.5673\big)$$

Credible Intervals

- If we dont know the functional form, but we can draw samples from the posterior, then we can use a Monte Carlo approximation to the posterior quantiles.

- We simply sort the S samples, and find the one that occurs at location α/S along the sorted list. As S→∞, this converges to the true quantile.

Run *mcQuantileDemo*
from Kevin Murphys' PMTK

Credible Intervals

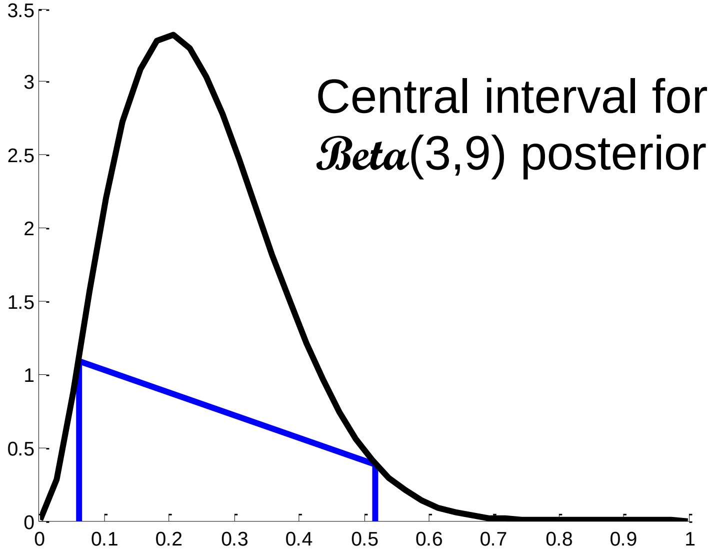
- Bayesian credible intervals versus frequentist confidence intervals are not the same thing.

- In general, credible intervals are usually what people want to compute, but confidence intervals are usually what they actually compute!

- Fortunately, the mechanics of computing a credible interval is just as easy as computing a confidence interval.

Highest Posterior Density Regions

- In central intervals there might be points outside the CI which have higher probability density. This is illustrated in the Figure, where we see that points outside the left-most CI boundary have higher density than those just inside the right-most CI boundary.



Central interval for a $\mathscr{Beta}$(3,9) posterior

Run *betaHPD*
from Kevin Murphys' PMTK

- This motivates the highest posterior density or HPD region. This is defined as *the set of most probable points that in total constitute* $100(1-\alpha)\%$ *of the probability mass*. More formally, we find the threshold p^* on the pdf such that

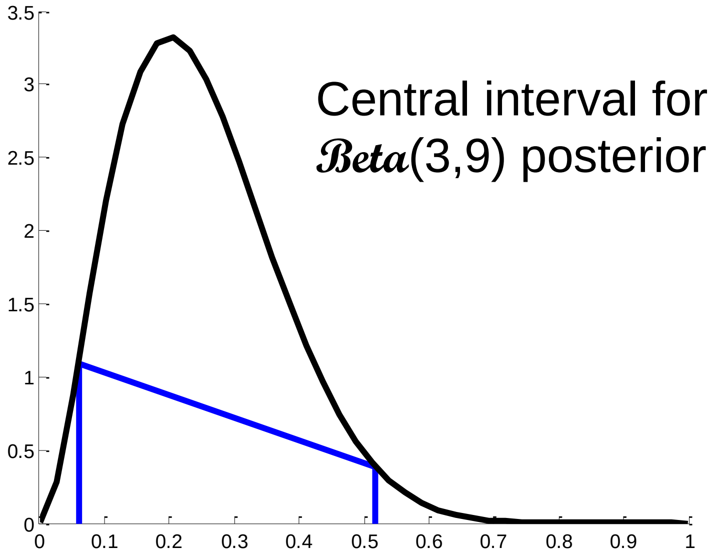
$$1-\alpha = \int_{\theta: p(\theta|\mathscr{D})>p^*} p(\theta|\mathscr{D})d\theta$$

- We then define the HPD as:

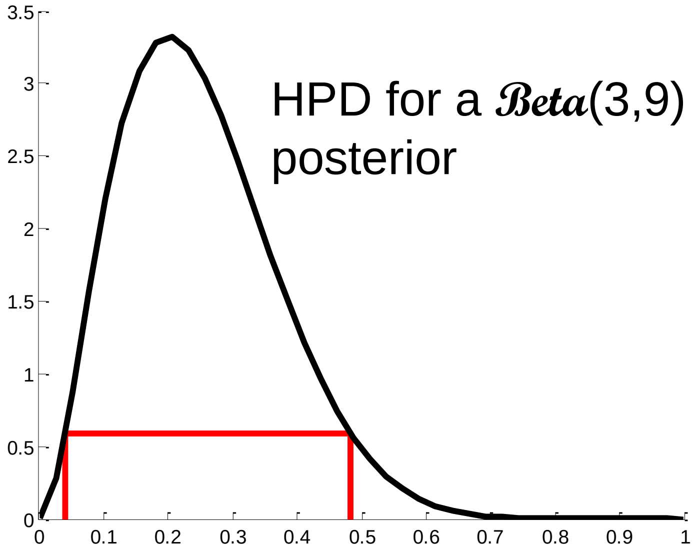
$$C_\alpha = \{\theta : p(\theta|\mathscr{D}) > p^*\}$$

Highest Posterior Density Regions

- *The HPD region is sometimes called a highest density interval or HDI*. The figure shows the 95%HDI of a $\mathcal{Beta}(3,9)$ distribution, which is $(0.04, 0.48)$.
- *We see that this is narrower than the CI*, even though it still contains 95% of the mass. Also *every point inside of it has higher density than every point outside of it.*

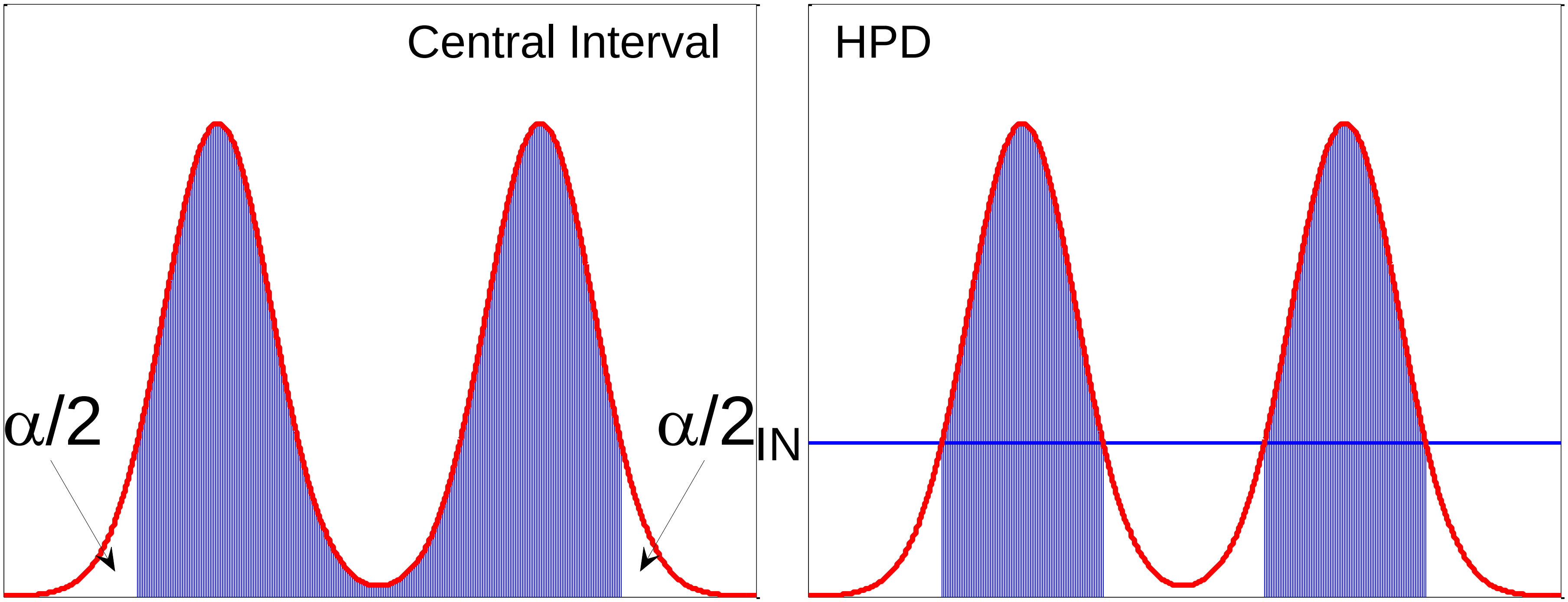

Central interval for a $\mathcal{Beta}(3,9)$ posterior

Run *betaHPD* from Kevin Murphys' PMTK


HPD for a $\mathcal{Beta}(3,9)$ posterior

- *The HPD region for unimodal distributions has* ***minimal width and contains*** 95% ***of the mass****. It can be computed by optimization and using the inverse CDF.*

Highest Posterior Density Regions



- For a unimodal distribution, the HDI will be the narrowest interval around the mode containing 95% of the mass.

- *If the posterior is multimodal, the HDI may not even be a connected region.* Note that summarizing multimodal posteriors is always difficult.

Run *postDensityIntervals*
from Kevin Murphys' PMTK

Inference for a Difference in Proportions

- Often we have multiple parameters, and we are interested in *computing the posterior distribution of some function of these parameters*.

- Example: suppose you are about to buy a book from Amazon.com

- Given:
  - a. Seller 1 has $y_1 = 90$ positive reviews and 10 negative reviews.
  - b. Seller 2 has $y_2 = 2$ positive reviews and 0 negative reviews.

- It seems you should pick seller 2, but we cannot be very confident that seller 2 is better since it has had so few reviews.

- We sketch a Bayesian analysis of this problem. Similar methodology can be used to compare rates or proportions across groups for a variety of other settings.

Inference for a Difference in Proportions

- Let θ_1 and θ_2 be the unknown reliabilities of the two sellers. We endow them both with uniform priors, $\theta_i \sim \mathcal{Beta}(1,1)$.

- The posteriors are $p(\theta_1 | \mathcal{D}_1) = \mathcal{Beta}(91,11)$ and $p(\theta_2 | \mathcal{D}_2) = \mathcal{Beta}(3,1)$.

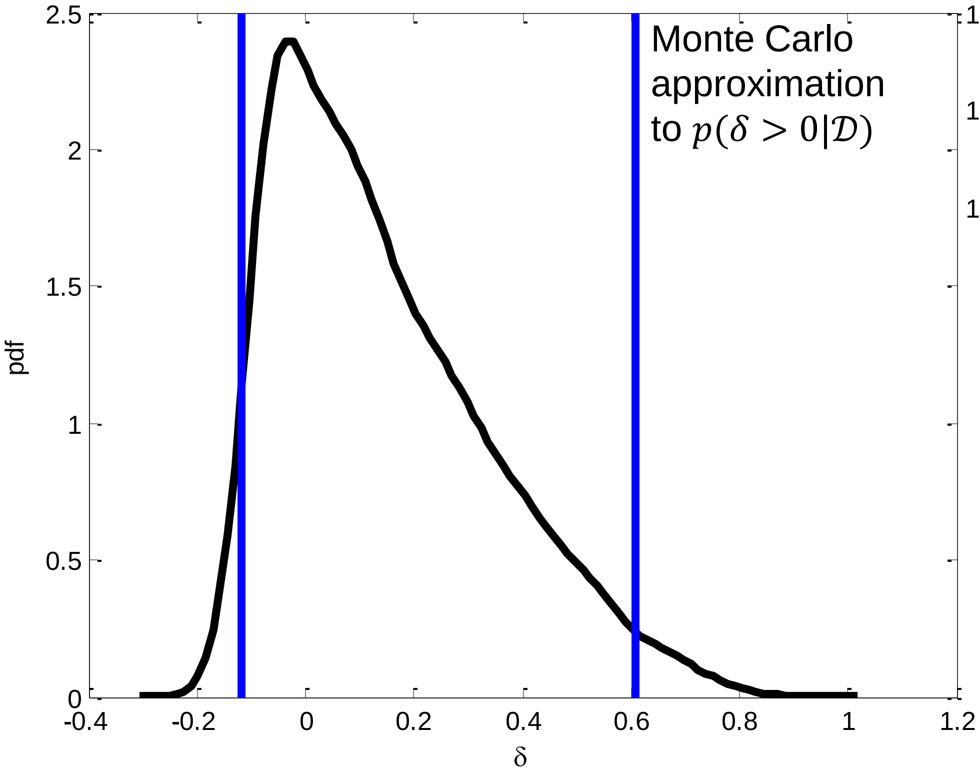
- We want to compute $p(\theta_1 > \theta_2 | \mathcal{D})$. For convenience, let us define $\delta = \theta_1 - \theta_2$ as the difference in the rates. We can compute the desired quantity using numerical integration:

$$p(\delta > 0|\mathcal{D})=\int_0^1 \int_0^1 \mathbb{I}(\theta_1 > \theta_2)\mathcal{Beta}(\theta_1|y_1+1, N_1 - y_1 + 1\big)\mathcal{Beta}(\theta_2|y_2+1, N_2 - y_2 + 1)d\theta_1 d\theta_2$$

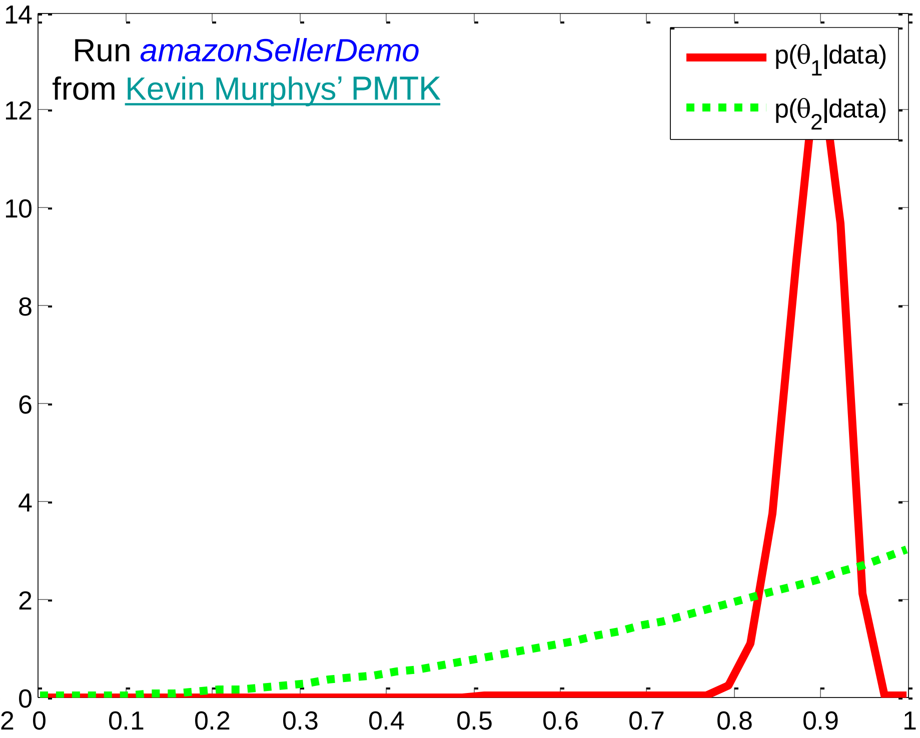
- We find $p(\delta > 0|\mathcal{D}) = 0.710$, which means you are better off buying from seller 1!

Run *amazonSellerDemo*
from Kevin Murphys' PMTK

Inference for a Difference in Proportions



Monte Carlo approximation to $p(\delta > 0|\mathcal{D})$



Run *amazonSellerDemo* from Kevin Murphys' PMTK

- We approximate the posterior $p(\delta|\mathcal{D})$ by MC sampling. θ_1 and θ_2 are independent and both have Beta distributions, which can be sampled easily.

- $p(\theta_i|\mathcal{D}_i)$ are shown on the right, and a MC approximation to $p(\delta|\mathcal{D})$ together with a 95% central interval on the left. An MC approximation to $p(\delta > 0|\mathcal{D})$ is obtained by counting the fraction of samples where $\theta_1 > \theta_2$. This turns out to be 0.718, which is very close to the exact value.

Model Selection

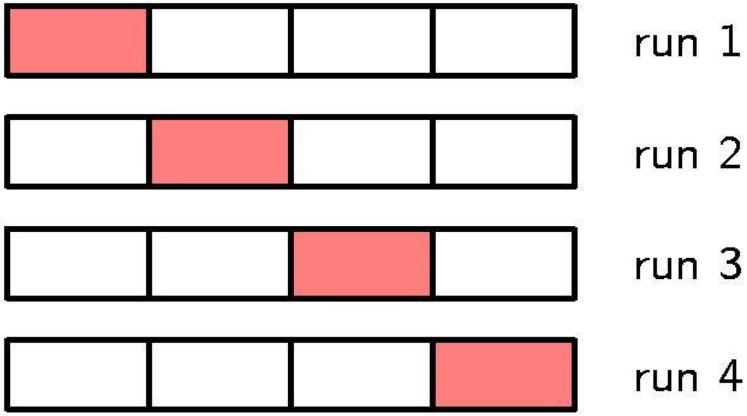
- ❑ A number of complexity parameters (polynomial order, regularization parameter, etc.) need to be selected to optimize performance/predictive capability. This is a model selection problem.

- ❑ In MLE, *the performance on the training set is not a good indicator of predictive performance due to the problem of over-fitting.*

- ❑ We often use some of the available data to train a range of models (or a given model with a range of values for its complexity parameters) and then to compare them on a validation set. We then select the one having the best predictive performance.

- ❑ Some over-fitting to the validation data can occur and a third test set on which the performance of the selected model is finally evaluated maybe needed.

Model Selection: Cross Validation



❑ The technique of S-fold cross-validation (here $S = 4$) involves taking the available data and partitioning it into S groups.

❑ $S - 1$ of the groups are used to train a set of models that are then evaluated on the remaining group. This procedure is repeated for all S possible choices for the held-out group and the performance scores from the S runs are then averaged.

Akaike Information Criterion

❑ The cross-validation cost increases by a factor of S.

❑ We should allow multiple hyperparameters and model types to be compared in a single training run.

❑ To correct for the bias of MLE, we use different information criteria (here M = # of parameters in the model), e.g.:

Akaike Information Criterion (AIC): $\ln p(\mathscr{D} \mid w_{ML}) - M$

We choose the model for which the AIC is largest.

❑ *AIC* does not account for uncertainty in model parameters. *It favor simple models.*

Bayesian Model Selection

- In general, when faced with a set of models (i.e., families of parametric distributions) of different complexity, how should we choose the best one?

  This is called the model selection problem.

- Examples:

  - *a low order polynomial in linear regression underfitts while a high order polynomial overfitts*

  - *a small regularization parameter λ results in overfitting and too large λ in underfitting*.

Bayesian Model Selection

- Can use CV to estimate the generalization error of all the candidate models, and then to pick the model that performs the best. This requires fitting each model K times, where K is the number of CV folds. More efficient approach is to *compute the posterior over models.*

$$p\left(m \mid \mathcal{D}\right) = \frac{p\left(\mathcal{D} \mid m\right) p(m)}{\sum_{m' \in M} p(m', \mathcal{D})}$$

- From this, we can easily compute the MAP model

$$\overline{m} = \max_{m} p\left(m \mid \mathcal{D}\right)$$

- This is called Bayesian model selection.

Model Evidence

- If we use a uniform prior over models, $p(m)\sim 1$, this amounts to picking the model which maximizes the marginal likelihood:

$$p(\mathscr{D}\,|\,m)=\int p(\mathscr{D}\,|\,\boldsymbol{\theta},m)\,p(\boldsymbol{\theta}\,|\,m)\,d\boldsymbol{\theta}$$

- This quantity is called the evidence for model m.

- The details on how to perform this integral will be discussed with examples later on.

- An intuitive interpretation of model evidence is discussed next.

Bayesian Occam's Razor

- ❑ One might think that using $p(\mathcal{D}|m)$ to select models would always favor the model with the most parameters.

- ❑ This is true if we use $p(\mathcal{D}|\hat{\theta}_m)$ to select models, where $\hat{\theta}_m$ is the MLE or MAP estimate of the parameters for model m - *models with more parameters will fit the data better, and hence achieve higher likelihood.*

- ❑ However, *if we integrate out the parameters, rather than maximizing them, we are automatically protected from overfitting*.

- ❑ Models with more parameters do not necessarily have higher marginal likelihood.

- ❑ This is called the Bayesian Occam's razor effect (MacKay 1995b; Murray and Ghahramani 2005)

- ❑ *Occams Razor Principle: one should pick the simplest model that adequately explains the data.*

Bayesian Occam's Razor

- The marginal likelihood can be rewritten as follows:

$$p(\mathcal{D}) = p(y_1)\,p(y_2 \mid y_1)\,p(y_3 \mid y_{1:2})\ldots p(y_N \mid y_{1:N-1})$$

where we have dropped the conditioning on m for brevity.

- This is similar to a leave-one-out cross-validation estimate of the likelihood, since we predict each future point given all the previous ones.

- If a model is too complex, it will overfit the early examples and will then predict the remaining ones poorly.

Bayesian Model Validation

- Suppose we have two models M_1 and M_2

- Each is associated with a set of parameters θ_1 and θ_2

- We consider priors $p_i(\theta_i \,|M_i)$, *likelihoods* $f_i(\boldsymbol{x} \,|\theta_i, M_i\,)$ and posteriors $p_i(\theta_i|\boldsymbol{x}, M_i\,)$

$$\pi_i(\theta_i \mid \mathbf{x}, M_i) = \frac{f_i(\mathbf{x} \mid \theta_i, M_i)\pi_i(\theta_i \mid M_i)}{\pi_i(\mathbf{x} \mid M_i)}$$

- We define as the *best* model the one that is more *probable to have generated* the data $\boldsymbol{x}$ that we observed.

Bayesian Model Validation

From data we can learn the parameters for each model and then the model itself

$$\boldsymbol{x} \Rightarrow \pi_i(\theta_i \mid \boldsymbol{x}, M_i) = \frac{f_i(\boldsymbol{x} \mid \theta_i, M_i)\pi_i(\theta_i \mid M_i)}{\pi_i(\boldsymbol{x} \mid M_i)} \Rightarrow \quad \pi_i(M_i \mid \boldsymbol{x}) = \frac{\pi_i(\boldsymbol{x} \mid M_i)\pi_i(M_i)}{\pi(\boldsymbol{x})}$$

Noting that

$$\pi_i(\boldsymbol{x} \mid M_i) = \int f_i(\boldsymbol{x} \mid \theta_i, M_i)\pi_i(\theta_i \mid M_i)d\theta_i$$

we can find the best model that represents the data by computing:

$$\frac{\pi(M_1 \mid \boldsymbol{x})}{\pi(M_2 \mid \boldsymbol{x})} = \frac{\pi(\boldsymbol{x} \mid M_1)\pi(M_1)}{\pi(\boldsymbol{x} \mid M_2)\pi(M_2)} = \underbrace{\frac{\int f_1(\boldsymbol{x} \mid \theta_1, M_1)\pi_1(\theta_1 \mid M_1)d\theta_1}{\int f_2(\boldsymbol{x} \mid \theta_2, M_2)\pi_2(\theta_2 \mid M_2)d\theta_2}}_{\substack{B_{10}^{\pi} \\ \text{Ratio of Bayes' factors}}} \underbrace{\frac{\pi(M_1)}{\pi(M_2)}}_{\substack{\text{Ratio of} \\ \text{Priors}}}$$

Bayesian Model Validation - Example

Consider the coin flipping example

Let θ the probability of getting heads

Consider two models:

M_1 Coin is Fair: $\theta|M_1 \sim \mathcal{B}(100{,}100)$
M_2 Coin is Unfair: $\theta|M_2 \sim \mathcal{B}(0.5{,}0.5)$

Data $\boldsymbol{x} = \{2H, 3T\}$

Bayes Factors $$\underbrace{\frac{\int f_1(\mathbf{x}|\theta, M_1)\pi_1(\theta|M_1)d\theta}{\int f_2(\mathbf{x}|\theta, M_2)\pi_2(\theta|M_2)d\theta}}_{\substack{\text{Ratio of} \\ \text{Bayes' factors}}} = \frac{\int \theta^2(1-\theta)^3\theta^{99}(1-\theta)^{99} \,/\, beta(100{,}100)d\theta}{\int \theta^2(1-\theta)^3\theta^{-0.5}(1-\theta)^{-0.5} \,/\, beta(0.5, 0, 5)d\theta} = \frac{0.031}{0.012}$$

Model Validation $$\frac{\pi(M_1|\mathbf{x})}{\pi(M_2|\mathbf{x})} = \underbrace{\frac{\int f_1(\mathbf{x}|\theta, M_1)\pi_1(\theta|M_1)d\theta}{\int f_2(\mathbf{x}|\theta, M_2)\pi_2(\theta|M_2)d\theta}}_{\substack{\text{Ratio of} \\ \text{Bayes' factors}}} \underbrace{\frac{\pi(M_1)}{\pi(M_2)}}_{\substack{\text{Ratio of} \\ \text{Priors}}} = 2.58\frac{\pi(M_1)}{\pi(M_2)}$$

Bayesian Model Validation - Example

Consider the coin flipping example

Let θ probability of getting heads

Two models:

M_1 Coin is Fair: $\theta|M_1 \sim \mathcal{B}(100,100)$
M_2 Coin is Unfair: $\theta|M_2 \sim \mathcal{B}(0.5,0.5)$

Data $\boldsymbol{x} = \{5H\}$

Bayes Factor
$$\underbrace{\frac{\int f_1(\boldsymbol{x}|\theta,M_1)\pi_1(\theta|M_1)d\theta}{\int f_2(\boldsymbol{x}|\theta,M_2)\pi_2(\theta|M_2)d\theta}}_{\substack{\textit{Ratio of}\\ \textit{Bayes' factors}}} = \frac{\int \theta^5\theta^{99}(1-\theta)^{99}\ /\ beta(100,100)d\theta}{\int \theta^5\theta^{-0.5}(1-\theta)^{-0.5}\ /\ beta(0.5,0,5)d\theta} = \frac{0.033}{0.25}$$

Model Validation
$$\frac{\pi(M_1|\boldsymbol{x})}{\pi(M_2|\boldsymbol{x})} = \underbrace{\frac{\int f_1(\boldsymbol{x}|\theta,M_1)\pi_1(\theta|M_1)d\theta}{\int f_2(\boldsymbol{x}|\theta,M_2)\pi_2(\theta|M_2)d\theta}}_{\substack{\textit{Ratio of}\\ \textit{Bayes' factors}}} \underbrace{\frac{\pi(M_1)}{\pi(M_2)}}_{\substack{\textit{Ratio of}\\ \textit{Priors}}} = 0.13\frac{\pi(M_1)}{\pi(M_2)}$$

Remark: Bayes’ factors and posterior model PDFs should be used with caution when non-informative priors are applied.

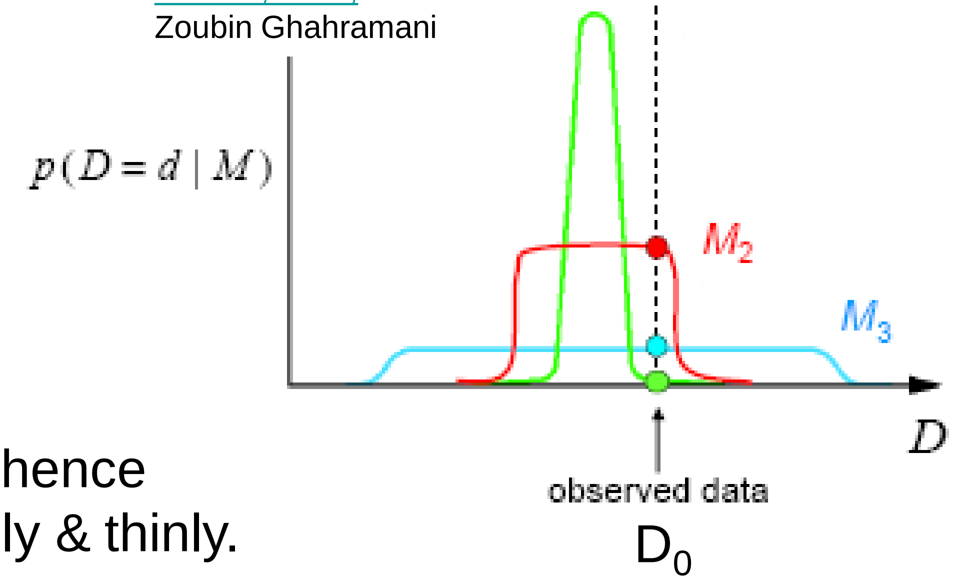
Bayesian Occam's Razor

- To further understand the Bayesian Occam's razor effect is to note that probabilities must sum to one (sum over all possible data sets)

$$\sum_{\mathscr{D}'} p\left(\mathscr{D}' \mid m\right) = 1$$

Bayesian Methods for Machine Learning, ICML Tutorial, 2004, Zoubin Ghahramani

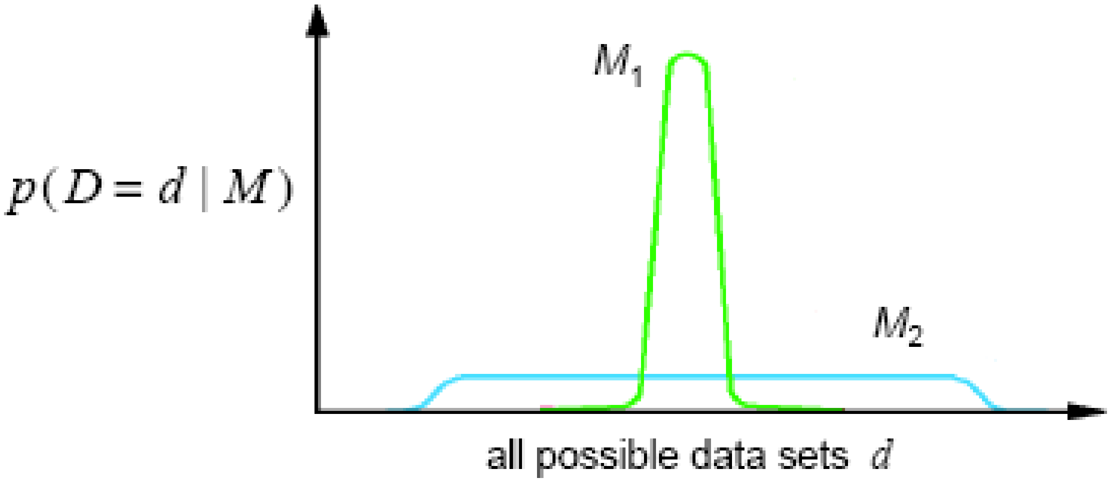
- Model 1 is too simple and assigns low probability to D_0.



- Model 3 also assigns D_0 relatively low probability, because it can predict many data sets, and hence it spreads its probability quite widely & thinly.

- Model 2 is "just right": it predicts the observed data with a reasonable degree of confidence, but does not predict too many other things. Hence model 2 is the most probable model.

Bayesian Occam's Razor



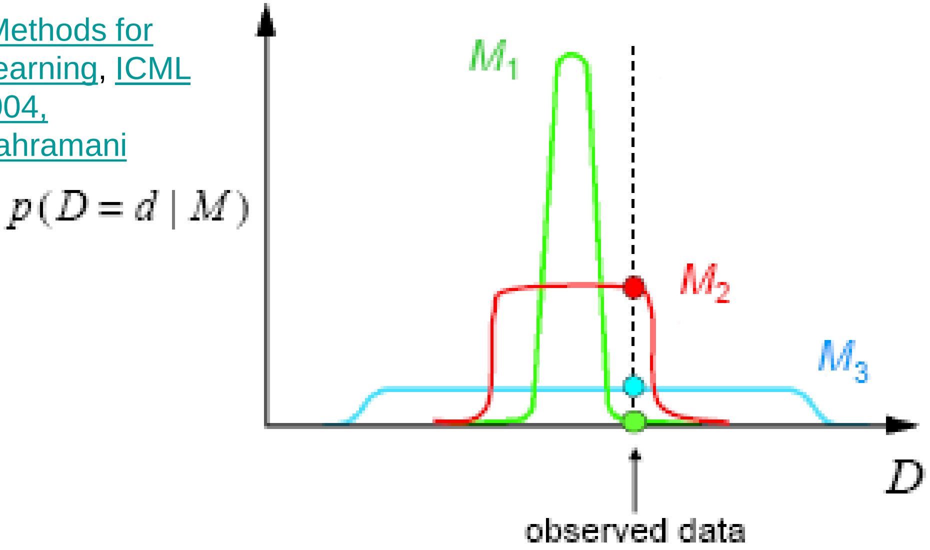
$$\text{For any model } M: \quad \sum_{\text{all } \mathbf{d}\in D} p(D=\mathbf{d}\mid M)=1$$

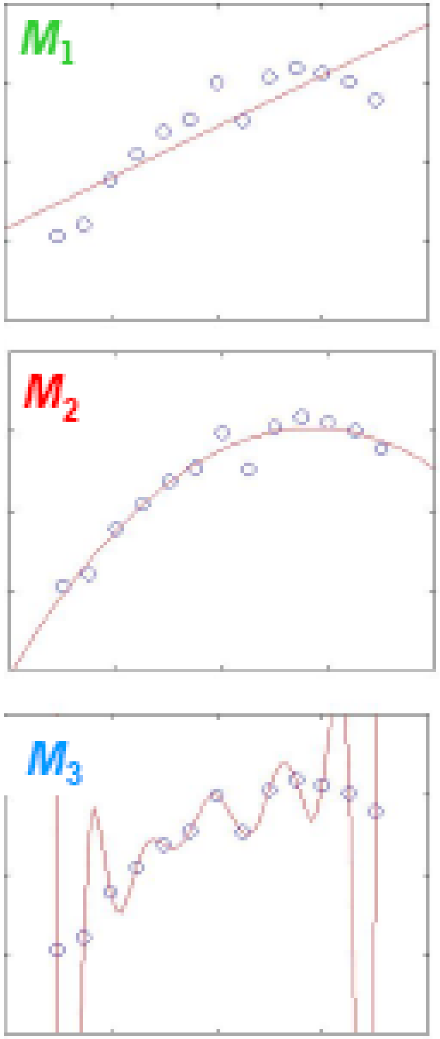
The law of *conservation of belief states that models that explain many* possible data sets must necessarily assign each of them a low probability

- A note on the evidence and Bayesian Occam's razor, I. Murray and Z. Ghahramani (2005), Gatsby Unit Technical Report GCNU-TR 2005-003
- Occam's Razor, C. Rasmussen and Z. Ghahramani, In T.K. Leen, T.G. Dietterich and V. Tresp (edts), Neural Information Procesing Systems 13, pp. 294-300, 2001, MIT Press

Bayesian Occam's Razor

Bayesian Methods for Machine Learning, ICML Tutorial, 2004, Zoubin Ghahramani





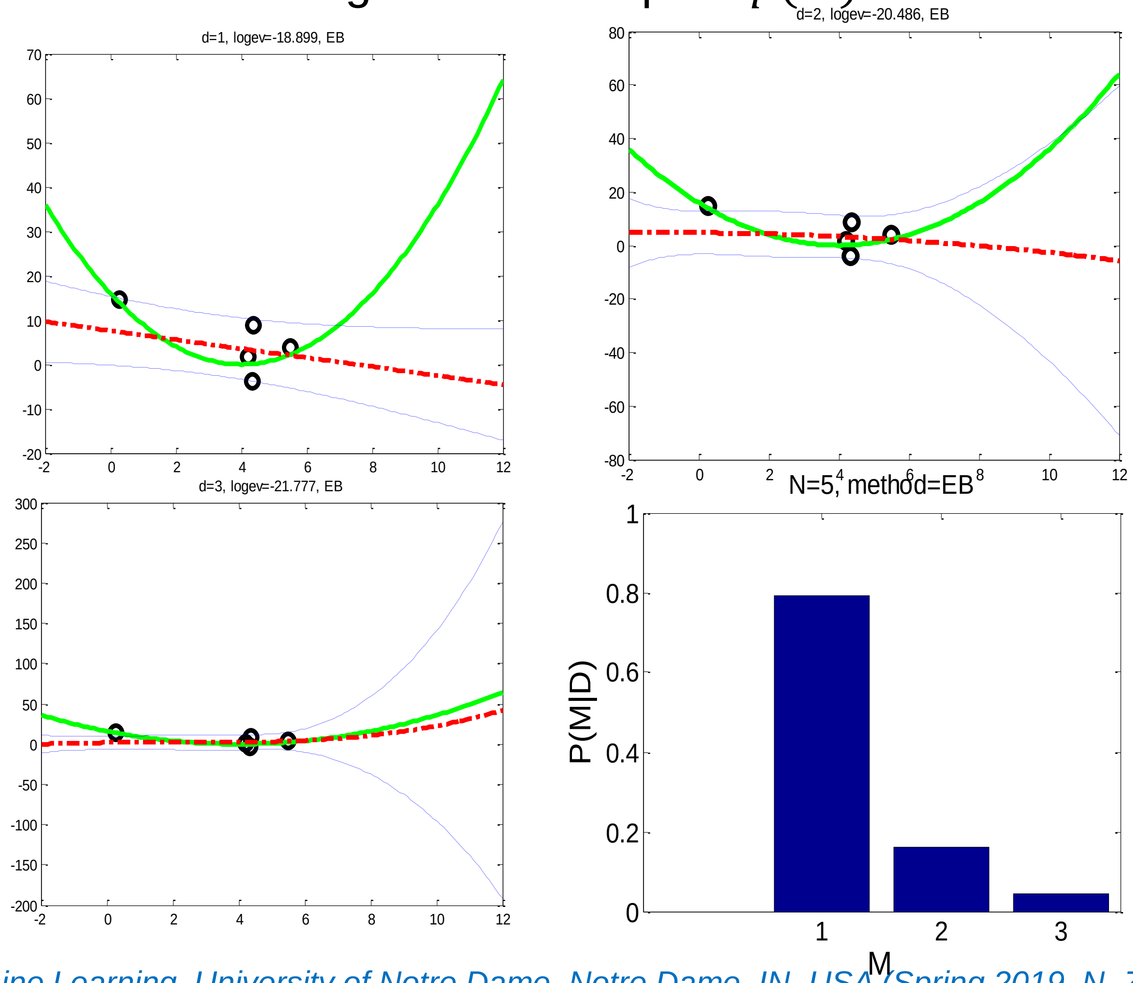
***M_1* : the too simple model is unlikely to generate this data**

***M_3* : the too complex model explains poorly a lots of data sets and it is a little better but still unlikely to have generated our data**

***M_2* : the just right model has the highest marginal likelihood**

Bayesian Occam's Razor

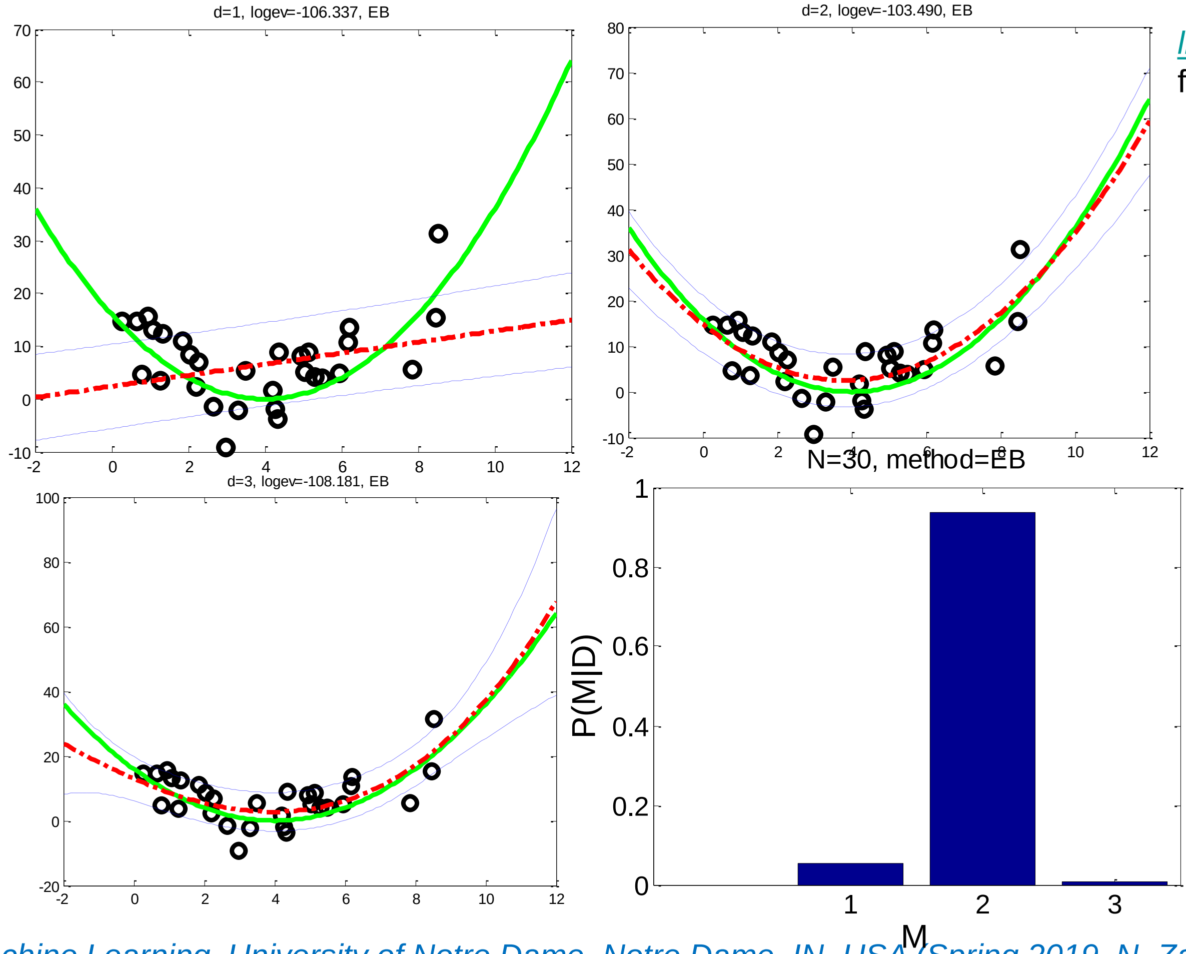
- Polynomials of degrees $1, 2, 3$ fit to $N = 5$ data points using empirical Bayes. Solid green curve is the true function, Dashed red curve is the prediction (dotted blue lines represent $\pm\sigma$ around the mean). The posterior over models $p(m|\mathcal{D})$ is also shown using a Gaussian prior $p(m)$.



linregEbModelSelVsN
from Kevin Murphys' PMTK

Bayesian Occam's Razor

- Polynomials of degrees $1, 2, 3$ fit to $N = 30$ data points using empirical Bayes. Solid green curve is the true function, Dashed red curve is the prediction (dotted blue lines represent $\pm\sigma$ around the mean). The posterior over models $p(m|\mathcal{D})$ is also shown using a Gaussian prior $p(m)$.



linregEbModelSelVsN
from Kevin Murphys' PMTK

Marginal Likelihood (Evidence)

- When discussing parameter inference for a fixed model, we often write

$$p(\theta \mid \mathscr{D}, m) \propto p(\theta \mid m)\, p(\mathscr{D} \mid \theta, m)$$

- We thus ignore the normalization constant $p(\mathcal{D}|m)$. This is valid since $p(\mathcal{D}|m)$ is constant wrt θ.

- However, when comparing models, we need to know how to compute the marginal likelihood, $p(\mathcal{D}|m)$.

- In general, this can be quite hard, since we have to integrate over all possible parameter values, but when we have a conjugate prior, it is easy to compute.

$$p(\mathscr{D} \mid m) = \int p(\mathscr{D} \mid \theta, m)\, p(\theta \mid m)\, d\theta$$

Marginal Likelihood - Evidence

- Let $p(\theta) = q(\theta)/Z_0$ be our prior, where $q(\theta)$ is an unnormalized distribution, and Z_0 is the normalization constant of the prior.

- Let $p(\mathcal{D}|\theta) = q(\mathcal{D}|\theta)/Z_l$ be the likelihood, where Z_l contains any constant factors in the likelihood.

- Let $p(\theta|\mathcal{D}) = q(\theta|\mathcal{D})/Z_N$ be our posterior, where $q(\theta|\mathcal{D}) = q(\mathcal{D}|\theta)q(\theta)$ is the unnormalized posterior, and Z_N is the normalization constant of the posterior.

$$p(\theta|\mathcal{D}) = \frac{p(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})} \Rightarrow \frac{q(\theta|\mathcal{D})}{Z_N} = \frac{q(\theta)q(\mathcal{D}|\theta)}{Z_0 Z_l p(\mathcal{D})} \Rightarrow p(\mathcal{D}) = \frac{Z_N}{Z_0 Z_l}$$

- So assuming the relevant normalization constants are tractable, we have an easy way to compute the marginal likelihood.

- Several examples are presented next.

Beta-Binomial Model

- ❑ Let us apply the above result to the Beta-binomial model. Since we know $p(\theta|D) = \mathcal{B}(\theta|a', b')$, where $a' = a + N_1$ and $b' = b + N_0$, we know the normalization constant of the posterior is $\mathcal{B}(a', b')$. Hence

$$p(\theta|\mathcal{D}) = \frac{p(\theta)\,p(\mathcal{D}|\theta)}{p(\mathcal{D})} = \frac{1}{p(\mathcal{D})}\left[\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\right]\left[\binom{N}{N_1}\theta^{N_1}(1-\theta)^{N_0}\right]$$

$$\frac{1}{B(a+N_1, b+N_0)} = \frac{1}{p(\mathcal{D})}\binom{N}{N_1}\frac{1}{B(a,b)}$$

$$p(\mathcal{D}) = \binom{N}{N_1}\frac{B(a+N_1, b+N_0)}{B(a,b)}$$

- ❑ The marginal likelihood for the Beta-Bernoulli model is the same as above, but without the $\binom{N}{N_1}$ term.

Dirichlet-Multinoulli Model

- One can show that the marginal likelihood for the Dirichlet-multinoulli model is given by

$$p\left(\mathscr{D}\right)=\frac{B(N+\alpha)}{B(\alpha)},\ B(\alpha)=\frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}$$

- Hence, we can rewrite the above result in the following form, which is more often used

$$p\left(\mathscr{D}\right)=\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\Gamma\left(N+\sum_{k=1}^{K}\alpha_k\right)}\prod_{k=1}^{K}\frac{\Gamma(N_k+\alpha_k)}{\Gamma(\alpha_k)}$$

Gaussian-Gaussian-Wishart Model

- Consider the case of a $\mathcal{MVN}$ with a conjugate $\mathcal{NIW}$ prior. Let Z_0 be the normalizer for the prior, Z_N be normalizer for the posterior, and let $Z_l = (2\pi)^{ND/2}$ be the normalizer for the likelihood. Then it is easy to see that

$$p(\mathscr{D}) = \frac{Z_N}{Z_0 Z_l} = \frac{1}{\pi^{ND/2}} \frac{1}{2^{ND/2}} \frac{\left(\frac{2\pi}{\kappa_N}\right)^{D/2} \left|S_N\right|^{-v_N/2} 2^{(v_0+N)D/2} \Gamma_D\left(v_N/2\right)}{\left(\frac{2\pi}{\kappa_0}\right)^{D/2} \left|S_0\right|^{-v_0/2} 2^{v_0 D/2} \Gamma_D\left(v_0/2\right)}$$

$$= \frac{1}{\pi^{ND/2}} \left(\frac{\kappa_0}{\kappa_N}\right)^{D/2} \frac{\left|S_0\right|^{v_0/2}}{\left|S_N\right|^{v_N/2}} \frac{\Gamma_D\left(v_N/2\right)}{\Gamma_D\left(v_0/2\right)}$$

- This equation will prove useful later on.

$$\mathcal{NIW}(\boldsymbol{\mu},\boldsymbol{\Sigma} \mid \boldsymbol{\mu}_0,\kappa_0,\boldsymbol{S}_0,v_0) = \mathcal{N}\left(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \frac{1}{\kappa_0}\boldsymbol{\Sigma}\right) \mathcal{InvWis}\left(\boldsymbol{\Sigma} \mid \boldsymbol{S}_0, \nu_0\right) =$$

$$= \frac{1}{Z_{NIW}} |\boldsymbol{\Sigma}|^{-1/2}\, exp\left(-\frac{\kappa_0}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right) |\boldsymbol{\Sigma}|^{-(\nu_0+D+1)/2} \exp\left(-\frac{1}{2} Tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_0\right)\right)$$

$$= \frac{1}{Z_{NIW}} exp\left(-\frac{\kappa_0}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) - \frac{1}{2} Tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_0\right)\right) |\boldsymbol{\Sigma}|^{-(\nu_0+D+2)/2}$$

$$Z_{NIW} = 2^{v_0 D/2} \Gamma_D\left(\frac{v_0}{2}\right) \left(\frac{2\pi}{\kappa_0}\right)^{D/2} \left|\boldsymbol{S}_0\right|^{-v_0/2}, \Gamma_D \text{ multivariate Gamma function}$$

Laplace Approximation

- The Laplace approximation allows a Gaussian approximation of the parameter posterior about the maximum a posteriori (MAP) parameter estimate.

- Consider a data set $\mathcal{D}$ and M models $\mathcal{M}_i, i = 1,..,M$ with corresponding parameters $\boldsymbol{\theta}_i, i = 1, \ldots M$. We compare models using the posteriors:

$$p(\mathcal{M} \mid \mathcal{D}) \propto p(\mathcal{M}) p(\mathcal{D} \mid \mathcal{M})$$

- For large sets of data $\mathcal{D}$ (relative to the model parameters), the parameter posterior is approximately Gaussian around the MAP estimate $\boldsymbol{\theta}_m^{MAP}$ (can also use 2^{nd} order Taylor expansion of the log-posterior):

$$p(\boldsymbol{\theta}_m \mid \mathcal{D}, \mathcal{M}_m) \approx (2\pi)^{-d/2} |\mathbf{A}|^{1/2} \exp\left(-\frac{1}{2} \left(\boldsymbol{\theta}_m - \boldsymbol{\theta}_m^{MAP} \right)^T \mathbf{A} \left(\boldsymbol{\theta}_m - \boldsymbol{\theta}_m^{MAP} \right) \right),$$

$$A_{ij} = -\left. \frac{\partial^2 \log P\left(\boldsymbol{\theta}_m \mid \mathcal{D}, \mathcal{M}_m \right)}{\partial \boldsymbol{\theta}_{mi} \partial \boldsymbol{\theta}_{mj}} \right|_{\boldsymbol{\theta}_m^{MAP}}$$

Laplace Approximation and Model Evidence

- ❑ We can write the model evidence as

$$p(\mathcal{D} \mid \mathcal{M}_m) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta}_m, \mathcal{M}_m)\, p(\boldsymbol{\theta}_m \mid \mathcal{M}_m)}{p(\boldsymbol{\theta}_m \mid \mathcal{D}, \mathcal{M}_m)}$$

- ❑ Using the Laplace approximation for the posterior of the parameters and evaluating the equation above at $\boldsymbol{\theta}_m^{MAP}$:

$$\log p(\mathcal{D} \mid \mathcal{M}_m)$$

$$\approx \log p(\mathcal{D} \mid \boldsymbol{\theta}_m^{MAP}, \mathcal{M}_m) + \log p(\boldsymbol{\theta}_m^{MAP} \mid \mathcal{M}_m) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{A}| + \frac{1}{2}\left(\boldsymbol{\theta}_m^{MAP} - \boldsymbol{\theta}_m^{MAP}\right)^T \boldsymbol{A}\left(\boldsymbol{\theta}_m^{MAP} - \boldsymbol{\theta}_m^{MAP}\right)$$

$$\approx \log p(\mathcal{D} \mid \boldsymbol{\theta}_m^{MAP}, \mathcal{M}_m) + \log p(\boldsymbol{\theta}_m^{MAP} \mid \mathcal{M}_m) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{A}|$$

- ❑ This Laplace approximation is used often for model comparison.

- ❑ Other approximations are also very useful:

  - Bayesian Information Criterion (BIC) (on the limit of $N \rightarrow \propto$)
  - MCMC (Sampling approach)
  - Variational Methods

Bayesian Information Criterion

- Start with the Laplace approximation for large data sets $N \rightarrow \propto$,

$$\log p(\mathcal{D} \mid \mathcal{M}_m) \approx \log p(\mathcal{D} \mid \boldsymbol{\theta}_m^{MAP}, \mathcal{M}_m) + \log p(\boldsymbol{\theta}_m^{MAP} \mid \mathcal{M}_m) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{A}|$$

- A N grows, $\boldsymbol{A}$ grows as $N\boldsymbol{A}_0$ for some fixed matrix $\boldsymbol{A}_0$, thus

$$\log|\boldsymbol{A}| \rightarrow \log|N\boldsymbol{A}_0| = \log\left(N^d |\boldsymbol{A}_0|\right) = d\log N + \log\left(|\boldsymbol{A}_0|\right) \xrightarrow{N\rightarrow\infty} d\log N$$

- Then the Laplace approximation is simplified as:

$$\log p(\mathcal{D} \mid \mathcal{M}_m) \approx \log p(\mathcal{D} \mid \boldsymbol{\theta}_m^{MAP}, \mathcal{M}_m) - \frac{d}{2}\log N \quad (as\ N \rightarrow \infty)$$

- Note interesting properties of (the easy to compute) BIC:
  - No dependence on the prior
  - One can use the MLE rather than the MAP estimate of (but use MAP when working with mixtures of Gaussians)
  - If not all parameters are well determined from the data, d =number of effective parameters.

- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461-464.

BIC Approximation to Log Marginal Likelihood

- The Bayesian information criterion or BIC thus has the following form:

$$BIC = \log p(\mathscr{D} \mid \overline{\boldsymbol{\theta}}_m, \mathscr{M}_m) - \frac{dof\left(\overline{\boldsymbol{\theta}}_m\right)}{2} \log N \approx \log p(\mathscr{D} \mid \mathscr{M}_m) \quad \text{(as } N \to \infty\text{)}$$

- $dof\left(\overline{\boldsymbol{\theta}}_m\right)$ is the number of degrees of freedom in the model, and $\overline{\boldsymbol{\theta}}_m$ is the MLE for the model. We see that this has the form of a penalized log likelihood, where the penalty term depends on the model complexity.

BIC for Linear Regression

- As an example consider linear regression. The MLE, log likelihood and BIC are:

$$MLE : \overline{\boldsymbol{w}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}, \quad \overline{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \overline{\boldsymbol{w}}^T \boldsymbol{x}_i\right)$$

$$\log p\left(\mathcal{D} \mid \overline{\boldsymbol{\theta}}\right) = -\frac{N}{2} \log\left(2\pi\overline{\sigma}^2\right) - \frac{\sum_i \left(y_i - \overline{\mu}\right)^2}{2\overline{\sigma}^2} \Rightarrow \log p(\mathcal{D} \mid \overline{\boldsymbol{\theta}}) = -\frac{N}{2} \log\left(2\pi\overline{\sigma}^2\right) - \frac{N}{2}$$

$$BIC = -\frac{N}{2} \log\left(2\pi\overline{\sigma}^2\right) - \frac{N}{2} - \frac{D}{2} \log N$$

- D is the number of variables in the model.

BIC for Linear Regression

- Hence the BIC score is as follows (dropping constant terms)

$$BIC = -\frac{N}{2}\log\left(2\pi\overline{\sigma}^{2}\right) - \frac{N}{2} - \frac{D}{2}\log N$$

- D is the number of variables in the model. In the statistics literature, it is common to use an alternative definition of BIC, which we call the BIC cost (since we want to minimize it):

$$BIC-Cost = -2\log p(\mathscr{D}\,|\,\overline{\boldsymbol{\theta}}_m, \mathscr{M}_m) + dof\left(\overline{\boldsymbol{\theta}}_m\right)\log N \approx -2\log p(\mathscr{D}\,|\,\mathscr{M}_m)$$

- In the context of the regression example, this becomes:

$$BIC-Cost = N\log\left(2\pi\overline{\sigma}^{2}\right) + N + D\log N$$

- The BIC method is related to the minimum description length or MDL principle. It characterizes the score of how well the model fits the data, minus how complex the model is.

Akaike Information Criterion

- There is a very similar expression to BIC/ MDL called the Akaike information criterion or AIC, defined as

$$AIC(m,\mathscr{D}) = \log p(\mathscr{D}|\bar{\boldsymbol{\theta}}_m, \mathscr{M}_m) - dof\left(\bar{\boldsymbol{\theta}}_m\right)$$

- This is derived from a frequentist framework, and cannot be interpreted as an approximation to the marginal likelihood.

- The penalty for AIC is less than for BIC.

$$BIC = \log p(\mathscr{D}|\bar{\boldsymbol{\theta}}_m, \mathscr{M}_m) - \frac{dof\left(\bar{\boldsymbol{\theta}}_m\right)}{2}\log N \approx \log p(\mathscr{D}|\mathscr{M}_m)$$

- This causes AIC to pick more complex models. However, this sometimes can result in better predictive accuracy!

- Clarke, B., E. Fokoue, and H. H. Zhang (2009). *Principles and Theory for Data Mining and Machine Learning*. Springer.

Effect of the Prior/Empirical Bayes

- When performing posterior inference, the prior may not matter too much since the likelihood often overwhelms the prior.

- But when computing the marginal likelihood, the prior plays a much more important role, since we are averaging the likelihood over all possible parameter settings, as weighted by the prior.

- If the prior is unknown, the correct Bayesian procedure is to put a prior on the prior. That is, we should put a prior on the hyper-parameter α as well as the $\boldsymbol{w}$. To compute the marginal likelihood, we should integrate out all unknowns, i.e., we should compute

$$p\left(\mathscr{D} \mid \mathbf{m}\right) = \iint p\left(\mathscr{D} \mid \mathbf{w}\right) p\left(\mathbf{w} \mid \alpha, m\right) p(\alpha \mid m) d\mathbf{w} d\alpha$$

Empirical Bayes

❑ This requires specifying the hyper-prior.

❑ Fortunately, the higher up we go in the Bayesian hierarchy, the less sensitive are the results to the prior settings. Thus can usually make the hyper-prior uninformative.

❑ A computational shortcut is to optimize a rather than integrating it out.

$$p(\mathscr{D}|\boldsymbol{m}) = \int p(\mathscr{D}|\mathbf{w})\, p\left(\mathbf{w}|\overline{\alpha}, m\right) d\mathbf{w}$$

where

$$\overline{\alpha} = \arg\max_{\alpha}\, p(\mathscr{D}|\alpha,\boldsymbol{m}) = \arg\max_{\alpha} \int p(\mathscr{D}|\mathbf{w})\, p(\mathbf{w}|\alpha, m)\, d\mathbf{w}$$

❑ This approach is called empirical Bayes (EB).

Back to Bayes Factors

❑ Suppose our prior on models is uniform, $p(m) \sim 1$. Then model selection is equivalent to picking the model with the highest marginal likelihood. Now suppose we just have two models we are considering, call them the null hypothesis, M_0, and the alternative hypothesis, M_1.

❑ Define the Bayes factor as the ratio of marginal likelihoods:

$$BF_{1,0} = \frac{p(\mathcal{D} \mid M_1)}{p(\mathcal{D} \mid M_0)} = \frac{p(M_1 \mid \mathcal{D})}{p(M_0 \mid \mathcal{D})} \Big/ \frac{p(M_1)}{p(M_0)}$$

❑ If $BF_{1\,0} > 1$, we prefer model 1, otherwise we prefer model 0. Jeffreys proposed a scale of evidence shown below

Bayes factor BF(1,0)	**Interpretation**
BF< 1/100	Decisive evidence for M_0
BF< 1/10	Strong evidence for M_0
1/10< BF< 1/3	Moderate evidence for M_0
1/3< BF < 1	Weak evidence for M_0
1 < BF < 3	Weak evidence for M_1
3 < BF < 10	Moderate evidence for M_1
BF>10	Strong evidence for M_1
BF>100	Decisive evidence for M_1

Bayes Model Selection: Jeffrey's Scale of Evidence

- Using the alternative reference below, Jeffrey's scale of evidence says:
  - □ For $\log\left(B_{10}^{\pi}\right)$ between 0 and 0.5, the evidence against H_0 is poor
  - □ In between 0.5 and 1, it is substantial
  - □ In between 1 and 2, it is strong and
  - □ Above 2, it is decisive.

$$B_{10}^{\pi} = \frac{\pi(x \mid H_1)}{\pi(x \mid H_0)}$$

- Bayes' factor tells us if one should prefer H_0 to H_1 (relative comparison of models).

- Bayes' factor does not tell us whether any of these models is sensible.

Estimation and Beyond in the Bayes Universe, Brani Vidakovic (online Course on Bayesian Stat. for Engineers)

Example: Testing if a Coin is Fair

- Suppose we observe some coin tosses, and want to decide if the data was generated by a fair coin, $\theta = 0.5$, or a potentially biased coin, where θ in $[0,1]$. Denote the fair coin model by M_0 and the biased coin model by M_1.

- The marginal likelihood under M_0 is simply

$$p(\mathscr{D} \mid M_0) = \left(\frac{1}{2}\right)^N$$

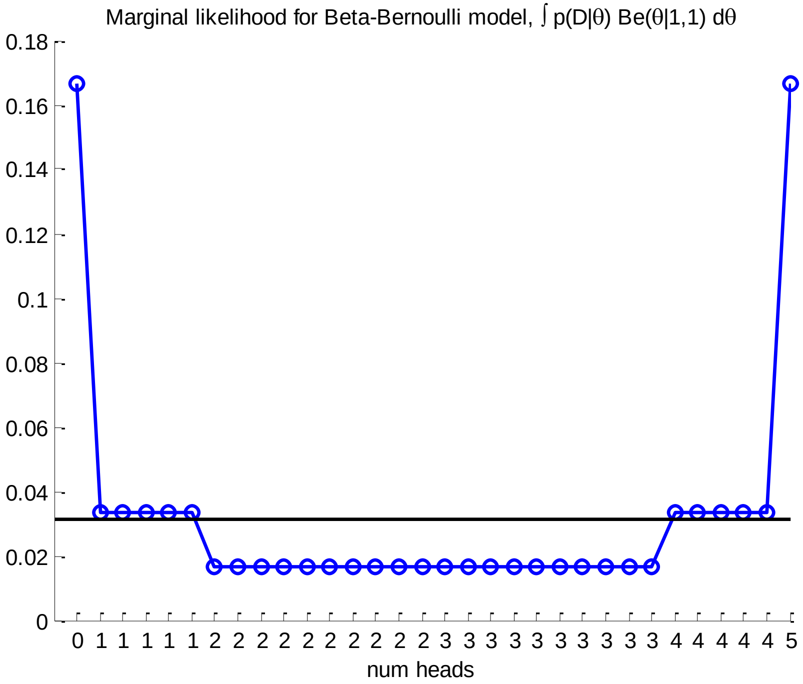
where N is the number of coin tosses.

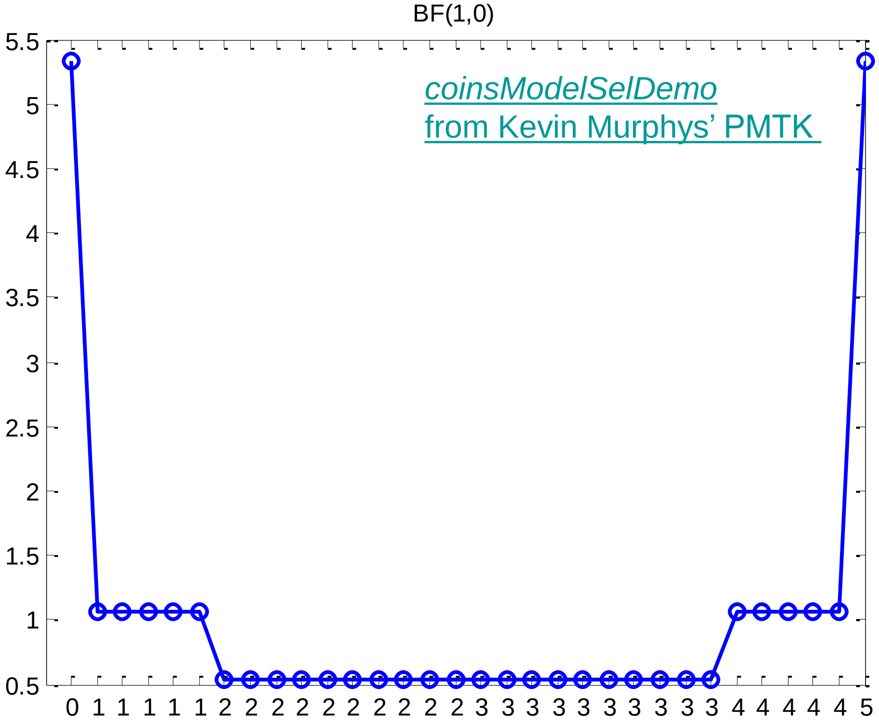
- The marginal likelihood under M$_1$ using a Beta prior, is

$$p(\mathscr{D} \mid M_1) = \int p(\mathscr{D} \mid \theta)\, p(\theta \mid M_1)\, d\theta = \frac{B(\alpha_1 + N_1, \alpha_0 + N_0)}{B(\alpha_1, \alpha_0)}$$

Example: Testing if a Coin is Fair

- ❑ We plot $\log p(\mathcal{D}|M_0)$ and $\log p(\mathcal{D}|M_1)$ vs the number of heads N_1 with $N = 5$ and $a_1 = a_0 = 1$.
- ❑ If we observe 2 or 3 heads, the unbiased coin hypothesis M_0 is more likely than M_1 since M_0 is a simpler model - it would be a suspicious coincidence if the coin were biased but happened to produce almost exactly $50/50$ heads/tails.
- ❑ However, as the counts become more extreme, we favor the biased coin hypothesis. Note that, if we plot the log Bayes factor, $logBF_{10}$ it will have exactly the same shape, since $logp(\mathcal{D}|M_0)$ is a constant.

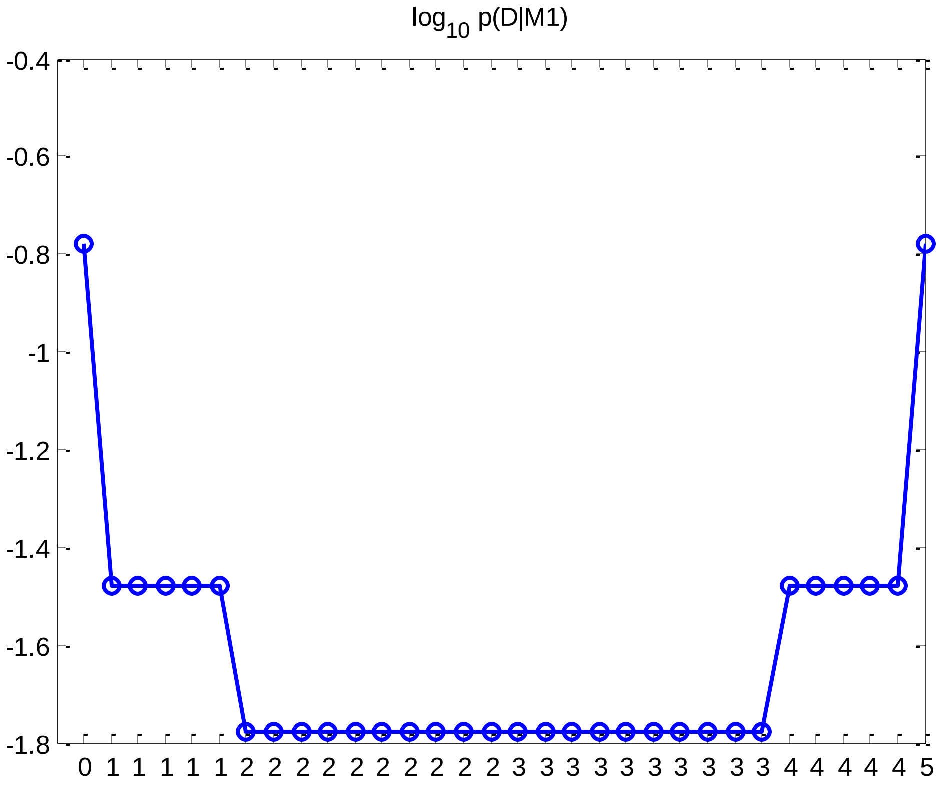




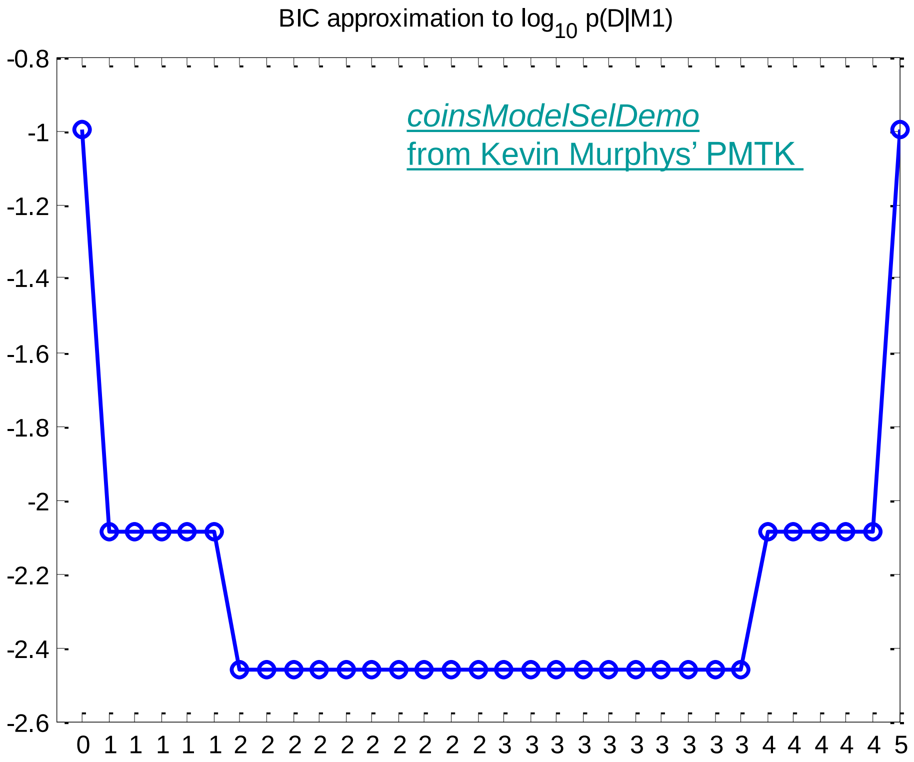
coinsModelSelDemo
from Kevin Murphys' PMTK

Example: Testing if a Coin is Fair

- Log marginal likelihood for coins example and the BIC approximation to $\log p(\mathcal{D}|M_1)$ for our biased coin example.
- The curve has approximately the same shape as the exact log marginal likelihood.
- It favors the simpler model unless the data is overwhelmingly in support of the more complex model.



log$_{10}$ p(D|M1)



BIC approximation to log$_{10}$ p(D|M1)

coinsModelSelDemo
from Kevin Murphys' PMTK

Jeffreys Lindley Paradox

- Define the marginal density of θ as: $p(\theta)=p(\theta|M_0)p(M_0)+p(\theta|M_1)p(M_1)$ where we consider the hypothesis $M_0:\theta\in M_0 \ vs \ M_1:\theta\in M_1$
- We can estimate the posterior as (denote: $p(M_0)=\rho,\ p(M_1)=1-\rho$)

$$p(M_0|\mathscr{D})=\frac{p(M_0)\,p(\mathscr{D}|M_0)}{p(M_0)\,p(\mathscr{D}|M_0)+p(M_1)\,p(\mathscr{D}|M_1)}=\frac{\rho\int_{\Theta_0}p(\mathscr{D}|\theta)p(\theta|M_0)d\theta}{\rho\int_{\Theta_0}p(\mathscr{D}|\theta)p(\theta|M_0)d\theta+(1-\rho)\int_{\Theta_1}p(\mathscr{D}|\theta)p(\theta|M_1)d\theta}$$

- Let us now assume that the priors are improper: $p(\theta|M_0)\propto c_0,\ p(\theta|M_1)\propto c_1$
- The posterior is determined by c_0/c_1 (e.g. it can be anything we want!)

$$p(M_0|\mathscr{D})=\frac{\rho\int_{\Theta_0}p(\mathscr{D}|\theta)d\theta}{\rho\int_{\Theta_0}p(\mathscr{D}|\theta)d\theta+(1-\rho)[c_1/c_0]\int_{\Theta_1}p(\mathscr{D}|\theta)d\theta}$$

- Using proper but vague priors causes similar problem.
- The Bayes factor favors the simpler model – complex models with diffuse priors have low probability.
- Jeffreys-Lindley paradox → Use proper priors for model selection. If M_0 & M_1 share the same prior over a subset of θ, this part of the prior can be improper, since c_0/c_1 will cancel out.